DOCUMENT RESUME

ED 431 802 TM 029 888

AUTHOR Kane, Michael
TITLE The Role of Generalizability in Validity.
PUB DATE 1999-04-00
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Error of Measurement; *Generalizability Theory; *Test Interpretation; *Validity
IDENTIFIERS *Invariance

ABSTRACT

The relationship between generalizability and validity is explained, making four important points. The first is that generalizability coefficients provide upper bounds on validity. The second point is that generalization is one step in most interpretive arguments, and therefore, generalizability is a necessary condition for the validity of these interpretations. It is also noted that the interpretive argument determines the appropriate estimate of generalizability. The fourth point is that generalizability provides justification for the syntactical content of construct labels. From each of these four perspectives, generalizability can be seen as a necessary, but not sufficient, condition for validity. The invariance assumptions are an integral part of validity, and standard errors and G coefficients quantify, in different ways, how well the invariance assumptions hold. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 802

THE ROLE OF GENERALIZABILITY IN VALIDITY

Michael Kane
University of Wisconsin
Madison, WI

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Michael Kane

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029888

BEST COPY AVAILABLE

Validity is the "bottom line" in evaluating the appropriateness of test-score interpretations, and provides a general framework for evaluating measurements. My thesis is that generalizability analyses play a central role in this framework.

I will examine the relationship between generalizability and validity from four points of view. First, generalizability coefficients provide upper bounds on validity. Second, generalization is one step in most interpretive arguments, and therefore, generalizability is a necessary condition for the validity of these interpretations. Third, the interpretive argument determines the appropriate estimate of generalizability. And fourth, generalizability provides justification for the syntactical content of construct labels.

1. Generalizability Provides an Upper Bound on Validity

Using classical test theory, we can show that criterion-related validity coefficients will be less than or equal to the square root of the reliability of the test scores. This statistical relationship between reliability and validity was initially derived for criterion-related validity coefficients, but it has been generalized to all cases. For example, if we want to estimate some theoretical construct for which we do not have a criterion measure, we can still assume that the construct has some definite value for each person and treat these values as a conceptual criterion. We can then imagine correlating our measurements with this conceptual criterion. In spite of the fact that we do not actually have the criterion, the statistical argument that the correlation of test scores with the conceptual criterion would be less than or equal to the square root of the reliability of the test scores still applies.

But clearly, the relationship between reliability and validity is more complicated than that presented in elementary textbooks. The core problem is that, in practice, we have many different

kinds of reliability coefficients, which may have very different values, and therefore different square roots. For example, in performance testing, it is not unusual to find inter-rater reliabilities in the .90s and internal consistency reliabilities in the .30s or .40s. But validity is taken to be a unitary concept. A test-score interpretation does not have multiple validities. So, given that we have multiple reliability estimates to choose from, the role of reliability in providing an upper bound on validity is, at best, a bit ambiguous.

The derivation showing that a reliability coefficients can provide an upper bound on validity also applies (with slight modifications) to generalizability coefficients. And since generalizability theory allows us to include all potential sources of error in a single coefficient, it helps to resolve the ambiguity created by multiple reliability coefficients.

The appropriate estimate of the G coefficient is not necessarily easy to define, depending as it does, on the design of the measurement procedure and the intended interpretation and use of the scores, but in most cases it is possible to define a G coefficient that includes all potential sources of error involved in using a measurement procedure for a particular purpose. For example, if the results of a performance test are to be used to estimate the general level of competence in an activity, we would probably want to generalize over tasks, raters, and occasions, plus residual variance, and therefore the G coefficient would reflect all of these sources of error. Ambiguity in the definition of the coefficient is less of a problem in generalizability theory than in classical test theory.

In general, the more detailed analysis of error in G theory will lead to larger estimated standard errors and lower coefficients than would result from classical reliability theory. However, in some cases, G theory may yield generalizability coefficients that are larger than traditional estimates of reliability. For example, if the test score is being used as an estimate of a

state variable, with the score obtained on a particular occasion being used to estimate the trait on that occasion, test-retest variability would not interfere with the inference. The measurement and the criterion could be perfectly correlated even if they both vary from occasion to occasion and therefore have low test-retest reliability. In fact, if the test score and the criterion vary in more-or-less the same way as a function of occasion, this dependency could enhance the test-criterion correlation as it lowers test-retest reliability.

In summary, the square root of an appropriately defined G coefficient constitutes an upper bound on validity. Generalizability theory tends to produce a more precise definition of the upper bound than is provided by classical reliability theory, because it provides a more thorough analysis of sources of error.

2. Generalization is a Critical Step in Most Interpretative Arguments

We can think of the interpretation assigned to test scores in any particular context in terms of a sequence of inferences leading from the score to conclusions and decisions. These inferences plus the assumptions on which they rest constitute the interpretive argument defining a proposed interpretation. The separate inferences in the argument must form an unbroken chain if the argument is to be considered valid. A failure of any link in the chain invalidates the argument as a whole. The interpretive argument provides an explicit and fairly detailed specification of the proposed interpretation, and therefore, provides guidance for validating the proposed interpretation. The validation is expected to provide adequate evidence for all of the inferences and assumptions in the argument.

Almost all interpretive arguments involve generalization as one major inference. A typical interpretive argument might start with a score based on examinee responses to some tasks

(e.g., test items, performance tasks, questionnaire items). It is assumed that the responses were collected under appropriate circumstances and scored using reasonable criteria, and therefore, that the score represents an accurate indication of the quality of the examinee's performance on those tasks under those conditions of observation.

In essentially all cases, this bare-bones interpretation does not get us where we want to go. So, the next step is to generalize the results to some universe of similar tasks or items. From the initial conclusion that the examinee did relatively well or badly on a particular set of tasks, we infer that the examinee can do this "kind of task" relatively well or badly. We also usually want to generalize over conditions of observation (raters, occasion, context, etc.). The inference from particular observations to a general conclusion about expected or typical performance over a universe of possible observations is an inductive inference. In G theory, this universe of possible observations is referred to as the universe of generalization, and the different kinds of conditions (e.g., tasks, raters, occasions) are referred to as facets. The interpretation typically involves additional inferences (e.g., to theoretical constructs, decisions), but my interest here is in generalization.

The justification for the inductive inference from the observed score to the universe score (the expected value over the universe of generalization) rests on certain law-like assumptions, which I will refer to as "invariance assumptions" or "invariance laws." An invariance assumption states that the results of applying the measurement procedure to a person (or other object of measurement) would not vary much if certain conditions of observation were changed. Invariance over raters assumes that the results would be approximately the same if we had employed a different set of qualified raters. Invariance over tasks assumes that the results would be about the same if we used another sample of tasks from the same pool of tasks. And, stability,

or invariance over occasions, suggests that the results would not vary much from one occasion to another.

Generalizability analyses provide empirical checks on these invariance assumptions. To the extent that the variance components associated with a facet (main effect and interactions) are small, the results are invariant over samples of conditions of this facet. In applying a measurement procedure, the required invariance assumptions may hold because there is very little variation over the conditions of a facet, or because the samples of conditions from the facet included in the measurement are fairly large.

To the extent that the overall standard error is small (and the corresponding G coefficient is large) it is reasonable to generalize over all facets included in the definition of error. On the other hand, a large value for the standard error indicates that one or more of the invariance assumptions has failed, and that the kind of generalization proposed in the interpretive argument is not justified. And as indicated earlier, if any inference in the chain fails, the argument as a whole fails. Therefore, an appropriate level of generalizability is a necessary condition for the interpretive argument to be plausible, that is, for the interpretation to be valid.

However, even if generalizability is very high, indicating that all of the invariance laws hold, and therefore that generalization is justified, some other link in the chain may fail. So, again, generalizability is not a sufficient condition for validity.

Note that poor generalizability can sometimes be corrected. Using a mutifaceted analysis of error, it is possible to identify those sources of error that are particularly large, and it may be possible to reduce these errors by adjusting the measurement procedure. For, example if the variance components involving raters tend to be particularly large, it may be possible to improve rater training or to increase the number of raters contributing to each observed score.

We can think of a failure of generalizability as invalidating one or more invariance assumptions, and therefore, as invalidating the interpretive argument.

3. The Proposed Interpretation Determines the Appropriate Generalizability Coefficient

Validation requires the development of evidence that justifies the proposed interpretation, or meaning, of scores. The scores must have an acceptable level of the right kind of generalizability for the interpretive argument to be plausible.

In this section, I will argue that the requirements inherent in validating a proposed interpretation determine the appropriate standard error and G coefficient.

As noted earlier, one problem with the use of reliability coefficients to set upper bounds on validity is the multiplicity of possible estimates of the reliability. There is no reason to expect different estimates of reliability(e.g., internal homogeneity, test-retest, interrater) to be even approximately equal, and there is no clear basis for choosing one of these estimates in preference to the others.

On the face of it, this situation is even worse in generalizability theory. The number of possible generalizability coefficients is much larger than the number of reliability coefficients in common use. We can choose to generalize over raters, or over occasions, or over tasks/items, or over any two of these facets or over all three. And of course, we can potentially include any additional facet we can imagine and from which we can draw a sample of conditions. Furthermore, we can consider facets to be fixed or random, or to involve sampling from a finite universe. And we can incorporate at least three different kinds of error (i.e., relative, absolute, regression-based). The total number of possible options is potentially quite large.

However, the definition of the standard error and of the corresponding generalizability coefficient are in fact tightly constrained by the proposed interpretation, and in particular, by the invariance assumptions associated with the proposed interpretations. The interpretive argument specifies the facets over which generalization is to occur and the range of conditions to be included in each facet and therefore the range of conditions over which generalization is to occur. Note that it is considerations about the plausibility of the overall interpretation (traditionally a validity issue) that are driving the definitions of the standard error and G coefficient.

The kind of error to be employed also depends on how the scores will be used. If the scores will, for example, be compared to some fixed standard of performance, an absolute error term is appropriate. If the scores will be used to make comparisons among individuals, a relative error is appropriate. Again, these are essentially validity issues.

The point here is that generalizability, like validity, depends on the proposed interpretation. If we interpret a score as a measure of an enduring trait (e.g., trait anxiety), we expect invariance over time, and therefore require that scores be generalizable over some extended period of time. If we interpret the score as a measure of a transitory, contingent state (e.g., state anxiety), we do not expect invariance over occasions, and therefore do not require generalizability over occasions. A trait interpretation assumes invariance over occasions and therefore requires evidence for invariance over occasions. A state interpretation treats the specific occasion on which the observation is made as a fixed part of the interpretation, and therefore does not require invariance over occasions.

The definition of an appropriate generalizability coefficient depends mainly on the proposed interpretation of the scores.

4. Generalizability Supports the Syntactical Content of Construct Labels

As noted above, generalizability is closely connected to the overall validity of a proposed interpretation in two reciprocal ways. Generalizability limits the interpretation, by limiting the extent to which the scores can be generalized; The interpretive argument can't assume invariance over a facet if scores are not invariant over the facet. In the other direction, the proposed interpretation determines the facets to be included in the standard error and the generalizability coefficient, and the way in which the facets will be defined (e.g.,broadly or narrowly).

The relationship of generalizability to validity can also be described in linguistic terms. The interpretation assigned to test scores can be thought of as the meaning, or semantic content, of the attribute label or construct label. Validation provides the supporting evidence for interpreting the scores, and thereby supports the proposed semantic content of the attribute label.

A part of the semantics of any term is its syntax, which specifies how the term is to be used. A noun can be used in one way; a verb is used in a different way. The attributes that we assign to objects of measurement function as adjectives, or more precisely using the terminology of first-order predicate logic, as predicates. The syntax of attributes, or predicates, depends mainly on the units to which they apply. Using the example mentioned earlier, trait anxiety can be represented as a one-place predicate, with persons as the sole argument:

$$A_T(p) = a_p$$

State anxiety is a two-place predicate, with persons and occasions as the two arguments:

$$A_S(p,o) = a_{po}$$

It is meaningful to talk about the trait anxiety of a person without specifying the occasion, but it is not meaningful to talk about state anxiety without specifying implicitly or explicitly, an occasion.

In either case, the observed score used to estimate the attribute depends on many conditions of observation. The world is much more complicated than our models of the world. The observed score can be represented as

$$X_{poIR\ldots}$$

Indicating the score obtained from person, p, on occasion, o, using items, I, raters, R, etc.

If we take $X_{poIR\ldots}$ as our estimate of $A_T(p)$, we are assuming that we can disregard any dependence on occasions, items, and raters, etc. If this assumption is to be viable, measures of trait anxiety must be largely invariant over these facets. In generalizing over certain facets, we are purposely simplifying our model of reality to make it more amenable to concise description and analysis. Generalization is over all facets other than persons, because the trait has been defined as a one-place predicate, with persons as the sole argument. This choice is part of the meaning of the attribute, the syntactical part of the overall interpretation.

For state variables, a different choice is made. State variables are defined as two-place predicates, which apply to person-occasion combinations. Therefore, for state variables, we do not generalize over occasions, because generalizing over occasions would not be consistent with the syntax of the attribute. A state variable is expected to change from one occasion to another, and these variations over occasions are part of the variability of interest. As a result, there is no reason to generalize over occasions, or to consider variability over occasions as a source of error.

The syntax of a term is a part of the meaning of the term. Validation justifies a particular interpretation (specified by the interpretive argument) for scores. Generalizability justifies the

invariance assumptions in the interpretive argument, and thereby, justifies the use of a certain syntax in describing scores. Again generalizability is a necessary but not sufficient condition for validity.

Overview

The four points that I have tried to make are essentially four ways of looking at the same relationship. Central to any interpretation are assumptions about the syntax of the attribute; these assumptions specify the kind of units to which the attribute applies, and by implication, the conditions of observation that are considered irrelevant to the interpretation. Most attributes apply to simple units (e.g., persons, or persons in particular settings), and their interpretive arguments involve generalization over all other conditions of observation. The invariance assumptions justifying this generalization define the sources of error to be included in the standard error and G coefficients, and these statistics, in turn, provide empirical checks on the invariance assumptions.

From each of these four perspectives, generalizability can be seen as a necessary but not sufficient condition for validity. The invariance assumptions are an integral part of validity, and standard errors and G coefficients quantify (in different ways) how well the invariance assumptions hold.